Analysis of finite capacity queuing networks

C. Osorio and M. Bierlaire

Transport and Mobility Laboratory, EPFL

Outline

- Project objectives
- Finite capacity queuing network
  - analysis methods
  - decomposition methods
- Proposed decomposition method
  - description
  - validation

Objectives

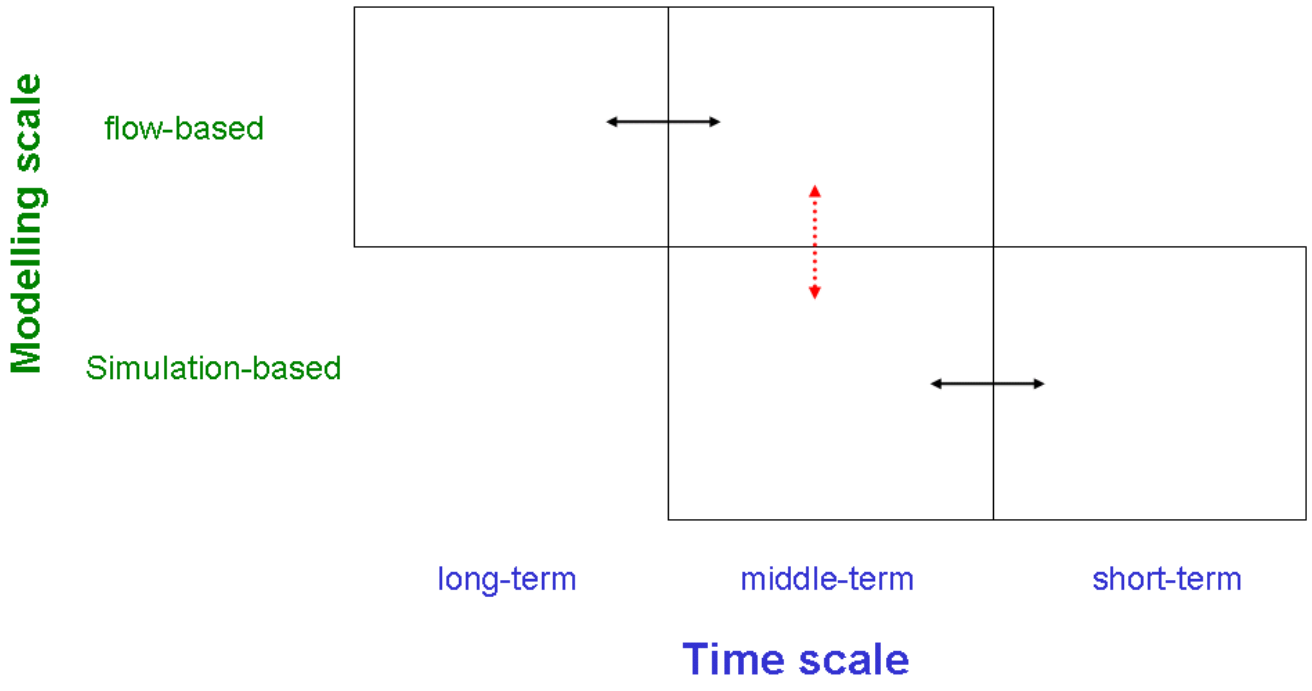
Context: hospital resource management.

2 main research tracks:

- macroscopic models:
  model aggregate flows: e.g. queuing theory.
  simpler to use and easier to integrate in an optimization process
- microscopic models:
  model specific details: simulation-based.
  more realistic model but cumbersome to optimize.

Long-term aim: define an optimization framework allowing the use of both approaches.

Objectives

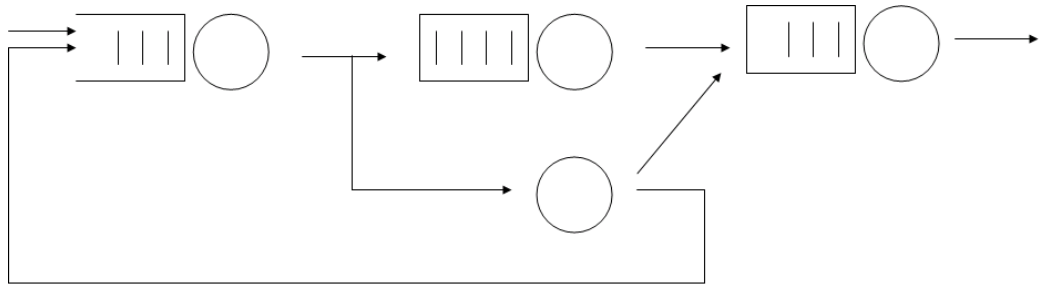


Current phase: define aggregate model using queuing theory

Finite capacity networks

- Jackson networks: infinite buffer size assumption violated in practice.
- Blocking may occur: **complex correlation structure** between the different queues in the network.

How can we model these networks?



Finite capacity queuing networks, FCQN

- General research:
  - Balsamo *et al*. 2001. Analysis of queuing networks with blocking.
  - Perros. 2001. Open queuing networks with blocking, a personal log.
  - Perros. 1984. Queuing networks with blocking: a bibliography.
- Field-specific research:
  - Balsamo *et al*. 2003. A review on queuing networks with finite capacity queues for software architectures performance prediction.
  - Artalejo *et al*. 1999. Accessible bibliography on retrial queues. Mathematical and computer modelling.
  - Papadopoulos *et al*. 1996. Queuing theory in manufacturing systems analysis and design : a classification of models for production and transfer lines.

FCQN methods

Aim: evaluate the main network performance measures using the joint stationary distribution, π.

1. Closed form expression
   exists only for a small set of networks:
   - product-form dbn: (Jackson, BCMP)
   - two-station single server with either tandem or closed topology

   For more general topology networks:
2. Exact numerical evaluation: solve $\pi Q = 0$
   Pb: construction of Q for the whole network: limited to small networks.
3. Approximation methods: decomposition methods

Decomposition methods

Aim: reduce dimensionality of the system under study by simplifying the correlation structure between the stations.

1. decompose the network into subnetworks
2. analyse each subnetwork independently: estimates of the marginal dbns
3. estimate the main performance measures

Analysis of a subnetwork:

i. use a network with a similar behaviour (e.g. Expansion method).

ii. analyse each subnetwork exactly and model their correlation via structural parameters.

Current objective

”Most existing blocking research takes either a tandem configuration with a single or multiple servers or an arbitrarily linked network model with feed-forward flows with a single server.” Koizumi (2005)

”No algorithms have been reported on networks of bufferless multiple server queues with the blocking-after-service rule.” Korporaal et al (2000)

Existing methods mainly concern:

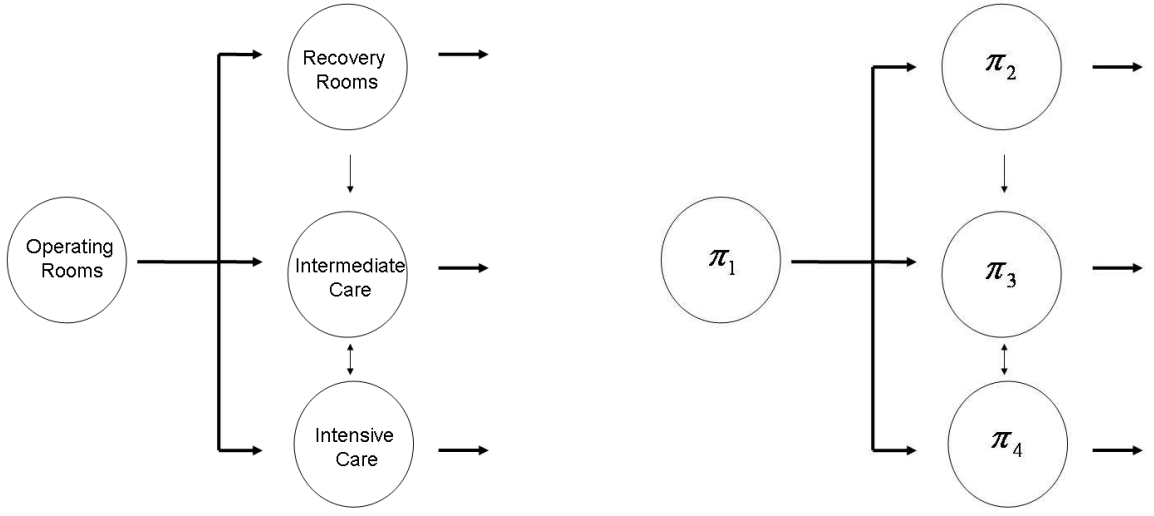
- single server queues in a feed-forward network
- multiple server queues in tandem

Current aim: **generalize to multiple server queuing networks with an arbitrary topology (allowing for feedback)**.

Decomposition method

Subnetwork size: single queues.

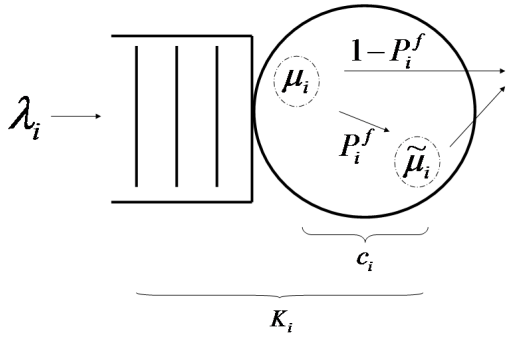
Aim: for each station i estimate the marginal distribution π_i.



This is done by solving the global balance equations. $\begin{cases} \pi_i Q_i = 0 \\ \sum_j \pi_{ij} = 1 \end{cases}$

Process description

Jobs go through an **active** phase, and may eventually go through a **blocked** phase:



- c_i parallel servers
- K_i total capacity: number of servers + buffer size
- λ_i, μ_i: average arrival and service rate
- $\tilde{\mu}_i$ average unblocking rate
- P_i^f average blocking probability

Transition rate estimations

Acknowledge correlation between stations: revise structural parameters.

Main challenge:

effective arrival rates
effective service rates

} stationary dbn of the subnetwork

$\updownarrow$

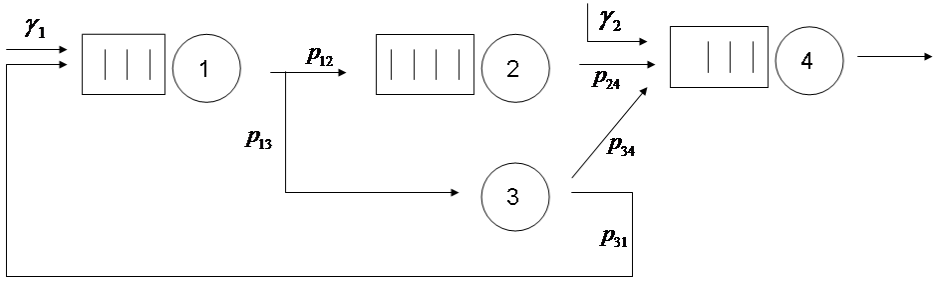
marginal stationary dbn of the network

$Q_i = f(\lambda_i, \mu_i, \tilde{\mu}_i)$

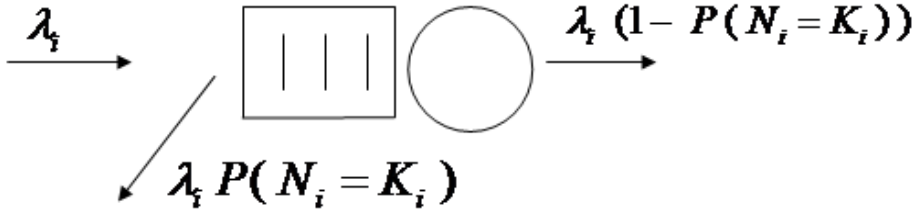
How can we estimate λ_i, μ_i and $\tilde{\mu}_i$?.

Arrival rates

Flow conservation laws: $\lambda_i^* = \gamma_i + \sum_{j \in i^-} p_{ji} \lambda_j^*$



Each station is modelled as a (two-dimensional) M/M/c/K queue, which are known as *loss models*:



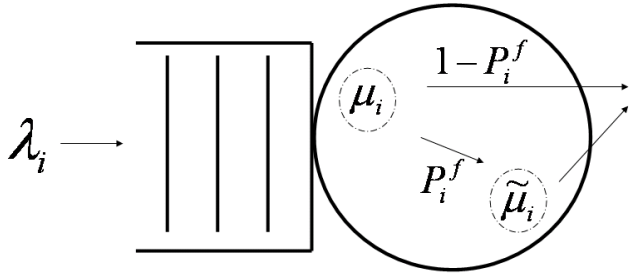
Arrival rates

The effective arrival rates are:

$$\lambda_i = \gamma_i + \sum_{j \in i^-} p_{ji} \lambda_j^* (1 - P(N_j = K_j))$$

Inter-arrival times: $T_i^A \sim \varepsilon(\lambda_i)$, iid
(i.e. Poisson arrival rates)

Service parameters



- active time: $T_i^A \sim \varepsilon(\mu_i)$, iid
- blocked time: $T_i^B \sim \varepsilon(\tilde{\mu}_i)$, iid

The average **effective** service time: $\frac{1}{\hat{\mu}_i} = \frac{1}{\mu_i} + P_i^f \frac{1}{\tilde{\mu}_i}$

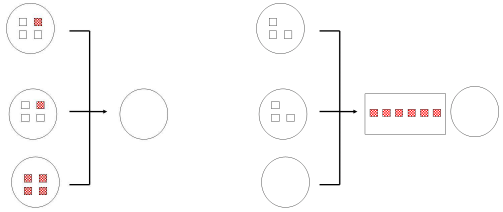
$P_i^f = \sum_{j \in i^+} p_{ij} P(N_j = K_j)$

How can we estimate the average blocked time $\frac{1}{\tilde{\mu}_i}$?

Service parameters

Blocked jobs can be seen as forming a virtual single server queue with a FIFO unblocking mechanism.

Aim: estimate the average waiting time in the virtual queue.



$$\frac{1}{\tilde{\mu}_i} = \sum_{j \in i^+} p_{ij} E[T_{ij}^B]$$

$$E[T_{ij}^B] \approx \frac{1}{2 r_{ij} \hat{\mu}_j c_j} (E[B_i] + 1) \frac{p_{ij} P(N_j = K_j)}{P_i^f}$$

where $r_{ij} = \frac{p_{ij} \lambda_i}{\lambda_j}$

Method validation

Comparing to:

- pre-existing decomposition methods
- exact solution on small networks.

Validation

Validation versus pre-existing methods

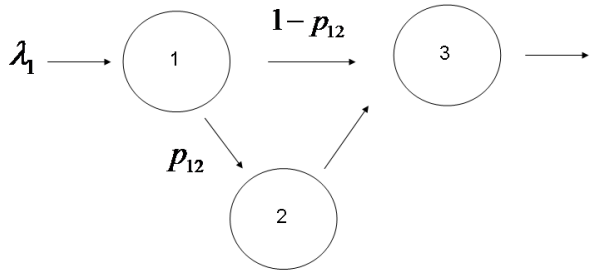
- Kerbache and MacGreggor Smith. 1988. Asymptotic behaviour of the Expansion method for open finite queuing networks. *Computers and Operations Research*
- Altiok and Perros. 1987. Approximate analysis of arbitrary configurations of open queuing networks with blocking. *Annals of Operations Research*
- Boxma and Konheim. 1981. Approximate Analysis of Exponential Queueing Systems with Blocking. *Acta Informatica*
- Takahashi *et al*. 1980. An approximation method for open restricted queuing networks. *Operations research*
- Hillier and Boling. 1967. Finite queues in series with exponential or erlang service times. A numerical approach. *Operations research*

Validation [1]

Setting: triangular topology with single-server stations ($c_j = 1$)

$\lambda_1 = 1, p_{12} = \frac{1}{2}$

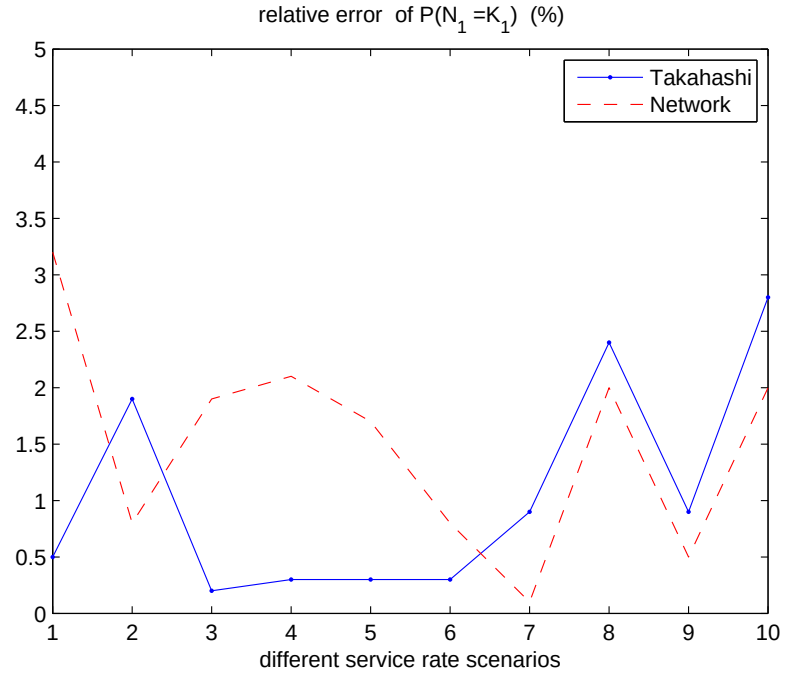
μ_1	μ_2	μ_3
1	1.1	1.2
1	1.2	1.4
1	1.3	1.6
1	1.4	1.8
1	1.5	2
1	1.6	2.2
1	1.7	2.4
1	1.8	2.6
1	1.9	2.8
1	2	3

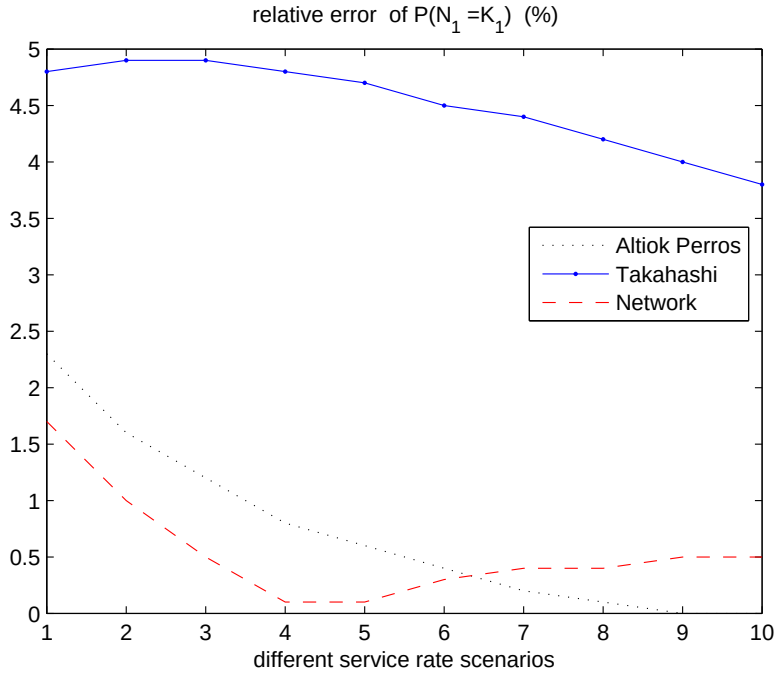


Validation [1]

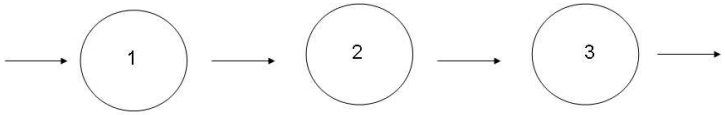
2 sets of scenarios:

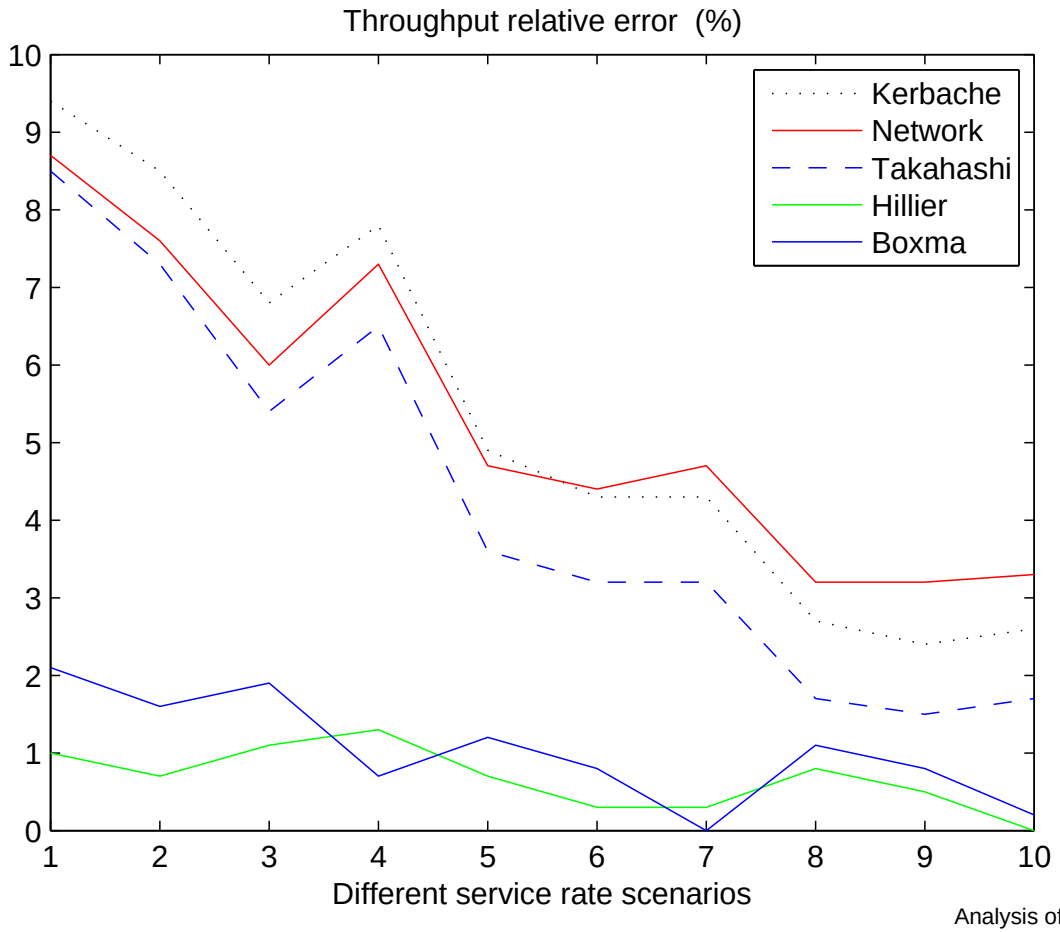
bufferless: $K_j = c_j = 1$, and non-bufferless: $K_j = 3$.





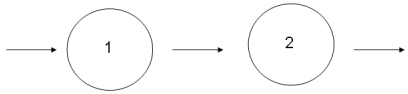
Validation [2]

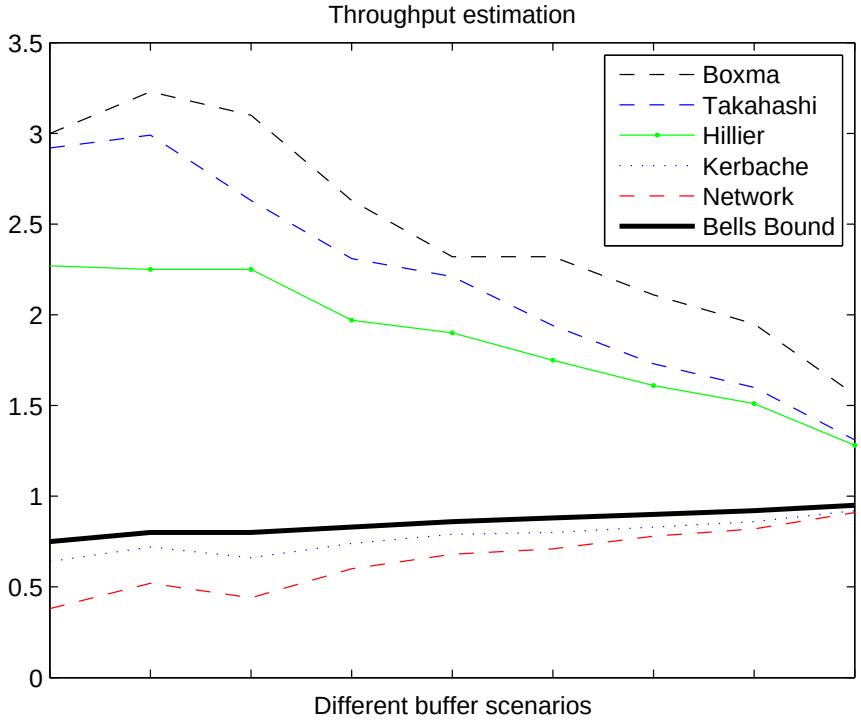




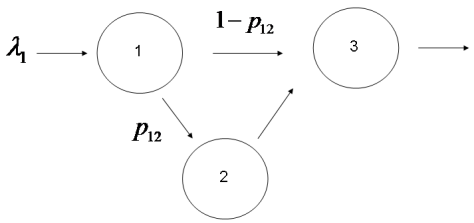
Validation [3]

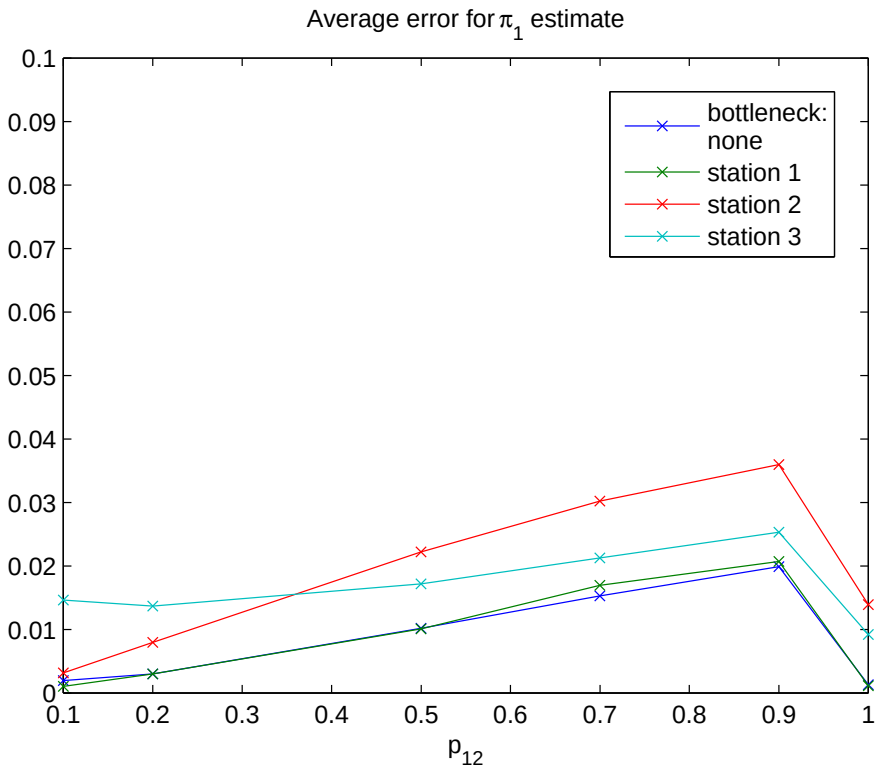
Theoretical bound on the throughput Bell (1982):





Validation versus exact dbn





Conclusions and future aims

Conclusions:

- a decomposition method allowing the analysis of finite capacity queuing networks has been proposed.
- its validation versus both pre-existing methods and exact models shows encouraging results
- this method explicitly models the blocking phase
- unlike pre-existing methods it preserves the original network topology and configuration (number of stations and their capacity)

Aims:

- Validation versus:
  - methods that account for networks with feedback.
  - simulation results on more complex networks.
- deadlock detection methods.